DOCUMENT RESUME

ED 404 709 EA 024 137

AUTHOR Huberty, Carl J.; Klein, Gerald A.
TITLE On Evaluating the Impact of an Innovative Educational Project.
PUB DATE Apr 92
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Data Collection; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Evaluation Problems; *Evaluation Research; *Program Effectiveness; *Research Methodology

ABSTRACT

The thesis of this paper is that, in evaluating most innovative educational projects that are conducted in schools, tenets of formal experimental design and associated inferential data analysis methods should be given limited emphasis. The basis of this thesis lies in the problems and difficulties that undermine the design and implementation of the typical project. Sections of the paper include discussions on design problems and difficulties, data collection, data analysis and reporting, statistical proportions as an evaluative measure, and graphs. These are followed by a presentation of the considerations that may be made in an evaluation report in the face of the problems and difficulties. (Contains 14 references.) (RR)

On Evaluating the Impact of an Innovative Educational Project

Carl J Huberty
University of Georgia

Gerald A. Klein
Georgia State Department of Education

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992.

On Evaluating the Impact of an Innovative Educational Project

Overview

The educational setting that will be considered herein is the public school; this includes elementary through high school and post-secondary school as well. The "subjects" may be students, teachers, administrators, programs, or even schools. A "project" effort may be developed for the purpose of enhancing the learning of students, of impacting on student attitude toward schooling, of decreasing the likelihood of student drop-out, of encouraging interdisciplinary teacher interaction, or of creating a more professional school atmosphere.

The intent of the evaluation activities to be discussed is on seeking evidence to support the conjectured effectiveness of the project of concern. The discussion presented in this paper will be restricted to the situation where the evidence collected for evaluation purposes is collected only on the project subjects. That is, the evaluation "design" to be considered will not include a so-called control or comparison school or school system.

Three examples of innovative educational projects are the following.

Example 1. This project provides an alternative educational plan to address students' individual needs without removing them from the regular school setting. The students selected are instructed in the academic areas of language arts, mathematics, social studies and science in a self-contained classroom setting. Curriculum has been modified to incorporate alternate strategies for at-risk students. In addition, eighth grade students are trained to work as tutors for at-risk third graders. Parents are encouraged to participate actively in their child's education through parent meetings and volunteer service.

Example 2. The purpose of this project is to provide opportunities for at-risk middle school students to improve overall academic achievement, self-esteem and attendance; to encourage parental involvement in education; and to improve decision-making, problem solving and cooperative skills among students. The provision of a positive education experience increases the likelihood that these students will stay in school and ultimately graduate from high school.

Example 3. The intent of this project is to improve teacher morale and enthusiasm by focusing on three aspects of wellness: personal, emotional and professional. Participation in a variety of activities allows teachers to demonstrate a) improved physical

and emotional wellness, b) reduced feelings of isolation from schoolwide and systemwide colleagues, c) increased awareness and perceptions of one's own professionalism, and d) increased morale and level of job satisfaction.

The thesis of this paper is: In evaluating most innovative educational projects that are conducted in the schools, tenets of formal experimental design and associated inferential data analysis methods should be given limited emphasis. The basis of this thesis lies in the problems and difficulties that undermine the design and implementation of the typical project. Some of these problems and difficulties are discussed in the second section. This will be followed by a presentation of considerations that may be made in an evaluation report in the face of the problems and difficulties.

Design Problems and Difficulties

The impact or effect of a typical project involves change or improvement with respect to some characteristic(s) of students, parents, teachers, administrators, classrooms, or schools. A first problem in assessing such effects is based on the measurement of the characteristic(s). Instruments that yield sensitive measurements with respect to the expected change may be difficult to identify or to develop. Even if such instruments are readily available, administration of them may be difficult because of scheduling or time, or both; testing takes up instructional time. Furthermore, testing only some selected students often requires parent permission, and may exceedingly disrupt class schedule. [School administrators and teachers are exhibiting considerable concern about "too much testing."] A second problem pertains to sampling of students, teachers, etc.. With most school-based projects, it is a we-take-what-we-can-get situation when it comes to sampling. Therefore, representativeness of a universe may be of some concern; thus there is potential for low external design validity. Also, if students are the sampling units of interest, there may very well be a potential for dependence of observations. The lack of something "close" to a probability sample plus the dependence of observations imply that formal statistical testing may very well be of questionable value. A third problem pertains to the lack of internal design validity. Often a design for the evaluation of a school project involves not only the units, usually students, who are directly involved in the innovation, but also some "control" or "comparison" group of units who were not directly involved in the innovative effort. What is often done is to compare the project group with the non-project group with respect to "average" performance level and then use such a comparison as the hard evidence of the existence of an effect. With such an in-the-field project, evaluators often want to attribute resultant differences (in favor of the project group, of course) to the innovation. We claim this is being fairly

naive. In a project that is implemented in a live school setting, the type of which we are assuming to be typical, there is little "control" over variables related to extraneous characteristics of units being studied that may affect the outcome(s). That is, if the project group exhibits more of an average gain than the non-project group, may the greater average gain be attributed to the project effort? Not necessarily!

We now turn to a discussion of some considerations in conducting an evaluation of an innovative project. In addition to addressing the three problems stated above, a focus will be given to a general approach to assessing project effects and to the reporting of results. Some exemplary results will be presented.

Data Collection

With specific regard to the measurement problem mentioned, above, it is recommended that as many measurements on the units or on objects affecting the units be obtained as is reasonable. Some measurements may be the traditional paper-and-pencil tests; others may be samples of day-to-day routine work, unobtrusive observations, formal interview responses, etc., etc.. Next for some specific suggestions regarding the problem of sampling. Rather than attempting to do some type of "representative sampling," one could focus on evaluating the exportability of the project (as opposed to being overly concerned about external design validity). Exportability pertains to the extent to which another school or school district can effectively adopt or adapt the project for a new setting. This quality of a project pertains more to applicability and adaptability as opposed to generalizability in the usual broad design and statistical sense. So then, the data collection effort should focus on those units -- children, teachers, or schools -- directly involved in the project.

Let us assume that the intent of the project is, in fact, one of affecting change. If so, it seems reasonable that past characteristics of the units -- relevant to the project effort, of course -- be obtained. How many months or years into the past one needs to pursue is a matter of judgment and resources. Some unit characteristics can be effectively changed in relatively little time, others must be attended to for some time. If the project, for example, is a three year effort, then it might be informative to have relevant information on comparable units for one or two pre-project years. This information, in addition to that collected during the three project years, may be useful in presenting evidence to support the project effort.

Numerical outcome measures are helpful as evidence to support the conjectured effectiveness of a project. Another type of information may, however, be helpful in providing supporting

evidence. Whereas the first-mentioned type of evidence is quantitative in nature, the second type is qualitative in nature. It is recognized that qualitative evaluation information may take various forms. The form we advocate pertains to information that describes project implementation as well as information that is outcome-related. Basically, what we have in mind are:

1. Student/teacher interviews;
2. Teacher activity logs;
3. Student/teacher questionnaires;
4. Classroom observation; and
5. Document analysis.

Data Analysis and Reporting

Considerations for data analysis in this paper are centered on quantitative data. The analysis of qualitative data and the reporting of results of such analyses are discussed by Patton (1990, chaps. 8 & 9), Strauss (1987, chap. 10), and Strauss and Corbin (1990).

The analysis of quantitative data and the reporting of results of such analyses are discussed by Fitz-Gibbon and Morris (1987), Morris, Fitz-Gibbon, and Freeman (1987), Popham (1988, chap. 10), and Wolf (1979, chap. 11). In these books attention is paid mostly to inferential methods; in particular, statistical testing. Most of the testing is devoted to the comparison of group means or adjusted group means. Little attention is paid to comparisons of any other score distribution characteristics. It is well known, for example, that outlying or extreme scores may have a considerable effect on a mean -- a discussion of this potential danger in an "inferential mean analysis" is not discussed in any of the above mentioned four books.

What will now be discussed are some alternatives to the group-mean-comparison analyses that typically involve formal statistical tests. It is assumed at the outset of this discussion that what is sought is some indication, on the basis of numerical data, of trend of student performance over time. A description of the progress or growth or change of performance of individual students over a period of some time would be extremely unwieldly. Furthermore, if we are to discuss an assessment of project effectiveness over years, the project would typically involve different groups of students. Thus, our discussion will mostly pertain to an investigation of progress of different groups of students for different years. A handful of different indices of change and different data presentation formats will be illustrated. The methods of analysis and reporting will not involve formal statistical tests. Rather, the focus is on informal comparisons in terms of: (1) proportions of students; and (2) graphical representations of some performances indicators.

The presentations given are based on real data collected for a project that involved high school students. Data were collected on different groups of students across three years -- one pre-project year and two project years (of a three-year project). A standardized achievement test (with four subtests) was administered at the end of Grade 9 and an equivalently-scaled test at the end of Grade 11. That is, for students in the Year 1 group we had a Grade 9 and Grade 11 score on each subtest; similarly for students in the Year 2 and Year 3 groups. The idea is to present test results in a way that will help (in addition to qualitative information) assess the effectiveness of the project. Such quantitative data might be analyzed using a simple analysis of variance on the 11th grade test scores or an analysis of covariance on the 11th grade test scores using 9th grade test scores as covariate scores. Rather, we suggest descriptive presentations of results. These presentations involve particular proportions of students in each year, and graphs of student performance index values.

Proportions

Various indices of group performance have been used in evaluation reports. A very common group performance index is the group mean; much less popular, but in some situations more appropriate than the mean, is the group median. An alternative index of group performance is a proportion of the group that performs above a given level. An example of such an index is the proportion of students who achieve at or above the midpoint of a five-point scale, at or above the 75th centile of a norm group for a standardized test, or at or above the median. The latter is used as an example; these proportions of students on the Grade 11 test are given in Table 1. From the table, it may be seen, for example, that 48.4% of the Year 2 students had Grade 11 Writing test scores that exceeded the norm group median. Based on these results given in Table 1, two conclusions might be drawn: (1) progress in Writing and Science is fairly evident and may reflect positive project effects; and (2) growth in Mathematics and Reading is positive from Year 1 to Year 2, but slight drops were realized from Year 2 to Year 3. The extent to which these results reflect support of the project effort is, of course, a judgment call. One thing that may be part of the judgment is project emphasis in the respective four academic areas during the first two project years. An evaluation report should include substantive explanations and support for any conclusions drawn.

A second index that might be considered for project evaluative purposes is the Grade 11 score minus the Grade 9 score on the test, or

Table 1

Proportions of Students Who Achieve at or Above The Norm Group Median on the Grade 11 Test

	Year1	Year2	Year3
Mathematics	45/96=.469	65/122=.533	68/141=.482
Reading	38/96=.396	65/122=.533	70/141=.496
Writing	41/96=.427	59/122=.484	80/141=.567
Science	47/96=.490	68/122=.557	81/141=.574

$$T_{11} - T_9 .$$

This is a simple "gain" score. A third index that may be favored by some evaluators takes into consideration the maximum possible gain that a student may attain:

$$\frac{T_{11} - T_9}{T_{max} - T_9} ,$$

where T_{max} is maximum attainable score on a Grade 11 test. Some evaluators may argue for the simple gain or difference score, $T_{11} - T_9$, while others would favor the percent or proportion of maximum possible gain as reflected in $(T_{11} - T_9)/(T_{max} - T_9)$. Which is preferable? A helpful selection suggestion has been made by Kaiser (1989): favor the one that has a lower correlation with T_9. For the exemplary data sets considered here, the mean absolute correlation (over the four subtests) for Year 3 is about .256 for T_9 versus $T_{11} - T_9$ and .087 for T_9 versus $(T_{11} - T_9)/(T_{max} - T_9)$. [Similar correlation patterns resulted for Year 1 and for Year 2.] Thus, in this case, $(T_{11} - T_9)/(T_{max} - T_9)$ is favored over the simple difference, $T_{11} - T_9$. To use the favored index for evaluation purposes, then, it was decided to examine the proportion of students who had an index score greater than some cut-off value. The cut-off value decided on was .20; there is nothing sacred or magical about this value. [Looking at the 12 (three years, four subtests) index distributions, .20 seemed to be a reasonable cut-off value.] The 12 proportions are given in Table 2. For example, it may be seen that 31.1% of the Year 1 students had a maximum possible percent of gain in Reading scores (from Grade 9 to Grade 11) that exceeded .20. Three conclusions from these results are: (1) the most impressive progress was in Science; (2) the Year 2 group exhibited a turn for the worse in Mathematics and Writing; and (3) progress in Reading was minimal.

Of course, these are judgment calls. Explanations for these conclusions and judgments would be included in the project evaluation report.

Graphs

As has been said, "A picture is worth a thousand words (numbers?)." Sometimes one is able to portray evaluation results more effectively if numerical summaries are supplemented with graphical summaries. Examples of graphs are the histogram, line chart, ploygon bar graph, pie chart, stemplot, and boxplot. Wolf (1979, pp. 174-175) and Morris et al (1987, chap. 3) present good

Table 2

Proportions of Students with a $(T_{11} - T_9)/(T_{max} - T_9)$ Value Greater than .20

	Year1	Year2	Year3
Mathematics	32/90=.356	33/116=.284	42/133=.316
Reading	28/90=.311	38/116=.328	46/133=.346
Writing	39/90=.433	29/115=.252	54/133=.406
Science	49/90=.544	67/115=.583	80/133=.602

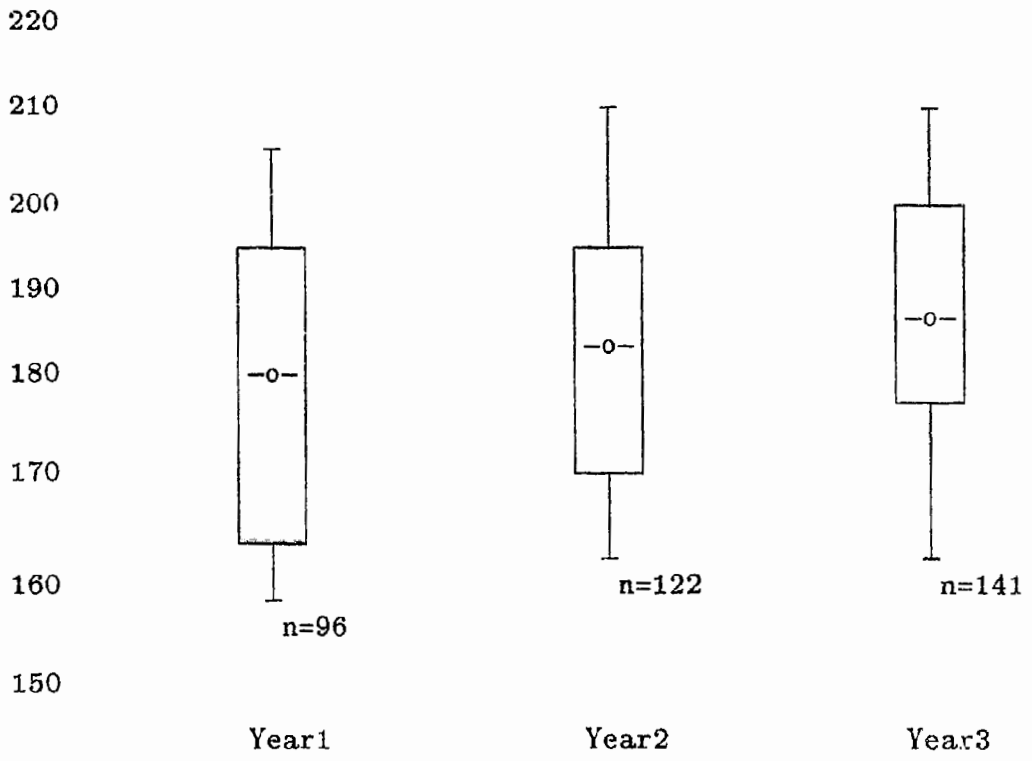
examples of bar graphs and line graphs. The current writers judge the stemplot and boxplot (see, e.g., Fitz-Gibbon & Morris, 1987, pp. 22, 31) to be particularly informative methods of representing characteristics of performance distributions. It is the boxplot that will be used to illustrate yet another means of representing evaluation information. The data sets alluded to earlier will again be used here -- Grade 9 and Grade 11 equivalently-scaled standardized tests (with four subsets) were administered to three successive project-year groups.

Whereas the numerical summaries presented in Table 1 and Table 2 involve comparisons of student scores with a cut-off value, the graphs will depict complete distributions. Two sets of boxplots will be presented. The index considered for the first plot is T_{11}, Grade 11 score; the Mathematics subtest score distribution was selected. The three boxplots are given in Figure 1. The upper edge of each box indicates the 75th centile, the lower edge indicates the 25th centile, and the line within the box indicates the 50th centile (i.e., the median). The mark

above the box indicates the 90th centile, and the mark below the box indicates the 10th centile.

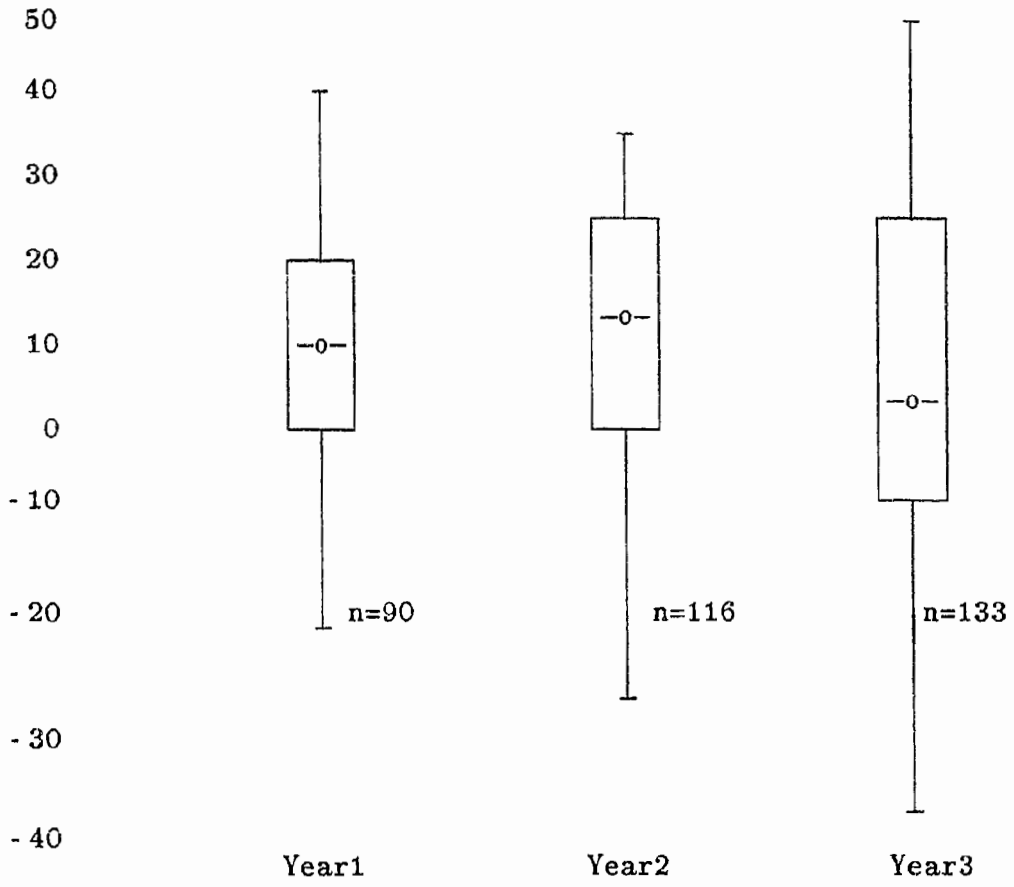
The plots in Figure 1 imply there was a nice steady growth in Grade 11 scores from Year 1 to Year 3. Whether or not the growth is "significant" is a judgment call. The gain from Year 1 to Year 3 across most distribution centiles is a little less than one standard error of measurement (which, as given in a test brochure, is about 8.0).

Figure 1. Boxplots of Grade 11 mathematics score distributions.



The index considered for the second plot is the proportion of maximum possible gain from Grade 9 to Grade 11, $(T_{11} - T_9)/(T_{max} - T_9)$, as discussed earlier in this paper. The three boxplots based on this index for the Reading subtest are given in Figure 2. [The respective group sizes given in Figure 2 are less than those in Figure 1 because for the proportion of maximum

Figure 2. Boxplots of $100(T_{11} - T_9)/(T_{max} - T_9)$ Reading score distributions.



possible gain, only students who took the test in Grade 9 and in Grade 11 were considered.] It is not clear from the information represented in Figure 2 that the percent of possible gain for students in Year 3 should be considered more than for students in Year 2 or in Year 1. Also, it may be noted that the distribution in Year 3 has greater dispersion than in the two prior years. It is fairly clear that information reflected in Figure 2 could not be used to support the conjectured effectiveness of the project. However, it should not be concluded from such a data display that, in fact, the project is ineffective, unless, of course, the objective of the project was to show "significant" or noteworthy increase across years with respect to this particular performance index.

Comments

It makes sense to the current writers that an evaluation of an innovative educational project should typically involve a combination of the use of qualitative-type and quantitative-type data. The discussion in this paper centered on the reporting of the quantitative-type data. Early in this paper it was argued that the oft-used traditional mean-based inferential analyses -- e.g., two-group t tests, ANOVA, ANCOVA -- may not be the most appropriate analysis of evaluation information. The argument against such analyses was based on questionable reliance on reasonable design internal and external validity. The usual conclusions drawn from mean-based analyses rely on such design validities, which we contend may be at a low level for in-the-field innovative educational projects. The data analysis methods suggested in this paper are descriptive in nature, and the reporting involved, simply, counts (or proportions or percents) and plots. It is recognized that more sophisticated approaches to the analysis of change have been presented (see, e.g., Embretson, 1991, and Willett, 1988).

The position taken herein is that the in-the-field educational community may very well be better served by evaluation reports on innovative educational projects that: (1) include the combination of qualitative-type and quantitative-type information; (2) focus on descriptive methods of analyzing and reporting the quantitative-type information; and (3) refer the descriptive evaluation results to results of related previously conducted research. The first two points were discussed in this paper. Following are a few words regarding the third point. A review of published research literature related to the project effort may yield some studies that consider contexts and variables similar to those of the project. Presumably, these studies were done in a fairly controlled setting; thus, it would be reasonable to conclude that the internal and external validity of the results were fairly high. Or, the related literature may simply include records or data on large groups of students or teachers similar to those involved in the project of concern. Reference may be made, then, to such research so that indirect (or, preliminary) project effects might imply project target effects.

Suppose, for example, published research revealed that about 90% of a certain type of ninth grade student will not graduate from high school; if a project has worked with a similar type of student and decreased the dropout rate by, say, 60%, a conclusion might be validly reached that the likelihood of high school graduation has been increased, even though truly conclusive data cannot be obtained until three years hence. For a second example, some research (e.g., Gladstone, 1987; Owens, 1988) indicates positive effects of interdisciplinary efforts of school teachers on learning by students. If a project evaluator can document effective implementation of interdisciplinary efforts,

then he/she might rely on relevant research for the support of project effectiveness. For a third example, some research (e.g., Gettinger, 1985; Leach & Tunnecliffe, 1984) suggests that increased student time-on-task leads to enhanced achievement. So, one type of evaluative information that suggests project effectiveness is the documentation of increased student time-on-task. The reasoning behind the use of previous research in support of project effectiveness may be presented as follows. The validity of the statement, "If A, then B," is supported by research in contexts comparable in some sense to that for the project of interest. If it may be established that "A" is confirmed in the context of the project, then "B" may logically be concluded.

The above reasoning would be more convincing, of course, if a number of published reports of empirical support for the effect in question (that is, the "B") could be cited. This is particularly the case if the contexts of the reported research are very similar to that of the local project of concern.

References

Embretson, S. E. (1991). Implications of a multidimensional latent trail model for measuring change. In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change (pp. 184-201). Washington, DC: APA.

Fitz-Gibbon, C. T., & Morris, L. L. (1987). How to analyze data. Beverly Hills, CA: Sage.

Gettinger, M. (1985). Time allocated and spent relative to time needed for learning as determinants of achievement. Journal of Educational Psychology, 77, 3-11.

Gladstone, C. (1987, November). Thinking, reading, and writing across the curriculum. Paper presented at the annual meeting of the National Council of Teachers of English, Los Angeles.

Kaiser, L. (1989). Adjusting for baseline: Change or percentage change? Statistics in Medicine, 8, 1183-1190.

Leach, D. J., & Tunnecliffe, M. R. (1984). The relative influence of time variables on primary mathematics achievement. Australian Journal of Education, 28, 126-131.

Morris, L. L., Fitz-Gibbon, C. T., & Freeman, M. E. (1987). How to communicate evaluation findings. Beverly Hills, CA: Sage.

Owens, T. R. (1988, April). Improving the collaboration of secondary vocational and academic educators. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Patton, M. Q. (1990). Qualitative evaluation and research methods. Newbury Park, CA: Sage.

Popham, W. J. (1988). Educational evaluation. Englewood Cliffs, NJ: Prentice Hall.

Strauss, A. L. (1987). Qualitative analysis for social scientists. New York: Cambridge.

Strauss, A., & Corbin, J. (1990). Basics of qualitative research. Newbury Park, CA: Sage.

Willett, J. B. (1988). Questions and answers in the measurement of change. Review of research in education, 15, 345-422.

Wolf, R. M. (1979). Evaluation in education. New York: Praeger.